IBM and EU partner to enable digitization of historic European texts on massive scale

August 25 2010

IBM and the EU have expanded their research collaboration, which now includes more than two-dozen national libraries, research institutes, universities, and companies across Europe to provide new technology that will enable highly-accurate digitization of rare and culturally significant historical texts on a massive scale.

Unlike past digitization projects where the result has been static, online libraries of texts, this unique widescale effort, called IMPACT (IMProving ACcess to Text), will offer new tools and best practices to institutions across Europe that will enable them to efficiently and accurately continue to produce quality digital replicas of historically significant texts and make them widely available, editable and searchable online.

Funded by the EU, IMPACT's research combines the power of new innovative Web-enabled adaptive optical character recognition (OCR) software with "crowd computing" technology -- a fast growing concept designed around individuals, or 'crowds,' enhancing a process or product by sharing their knowledge and expertise to dramatically improve its quality and efficiency. Combined, these technologies will allow institutions for the first time to adapt digitization to the idiosyncrasies of old fonts, anomalies and even vocabularies-while reducing error rates by 35% and substitution rates by 75%.

"IMPACT is remarkable in that it not only allows these prominent centers of culture to ultimately bring people closer to perhaps never

before seen historically significant texts of heritage -- but because it actually allows these people to become part of the preservation process," said Tal Drory, manager of the document processing group at IBM Research in Haifa. "IMPACT offers the first digitization system that combines the power of crowd computing with an adaptive optical character recognition (OCR) correction solution that can achieve excellent recognition rates across all kinds of documents - from the 15th century right up through the 19th century."

Rescued from fire and water at the Bavarian State Library in March of 1943, Karl von Eckartshausen's Magic: Principles of Higher Knowledge is an example of one special piece of work being digitized through IMPACT technology.

While today's OCR engines perform well with modern printed texts, the faded ink, age and unusual shapes of older typefaces can lower recognition rates by up to 50% and require massive manual post-production review. Consequently, for large-scale projects such as this, the efficiency of post-production review of digitized text is crucial. "The only way to make a large-scale digitization project work is to dramatically improve the quality of the initial OCR, and cut down post-processing tasks as much as possible," said Hildelies Balk, Head of European Projects at Koninklijke Bibliotheek and leader for the IMPACT consortium. "With IMPACT, we're expecting to see remarkable increases in productivity in the digitization process."

At the core of the digitization project lies a new, unique collaborative correction system, designed by IBM researchers, that makes it simple and convenient for large groups of volunteers spread over the continent to verify the accuracy of processed texts and correct recognition mistakes using an online web system. Moreover, inherent in the system is the ability to learn from its recognition errors, and adapt automatically to the specific font's characters.

IMPACT technology streamlines, simplifies and accelerates the process of winnowing out questionable text scans, enabling reviewers to key in corrections to the text. Instead of displaying an entire scanned page, reviewers only see the actual letters or words in question. For example, the letter combination "r" and "n" ("rn") may appear indistinguishable from the letter "m." In those instances, the system collects many instances of the letter "m," and places these samples next to the letters in question, making it much easier to determine the letter's real identity.

In cases where an entire word is suspect, it is added to a collection of other questionable terms, which are then arranged in alphabetical order. Volunteer reviewers need only accept or reject suggested substitutes with one keystroke. In addition, the system uses adaptive dictionary enrichment, a method in which new words are added to a central dictionary based on cross-identification and correction by other users.

For example, a small book that normally takes four hours to key in manually, would take one hour using standard OCR technology with manual correction. Incorporating the new collaborative review technology cuts the process down to 30 minutes. IBM researchers explained that the new adaptive OCR system can further reduce the time, cutting it in half to 15 minutes.

IBM Haifa researchers have experience in developing unique approaches to OCR that have proved themselves over the years, from tools that help categorize, classify and route mail and packages in large postal systems, to solutions for optimizing the reading of license plates in congestion pricing systems. IMPACT is likely the first real attempt to develop an adaptive OCR engine that is specifically designed for digital library purposes.

Source: IBM